German Council for Social
and Economic Data (RatSWD)

RatSWD

Working Paper Series

Working Paper No. 99

Regional Data

Gerd Groezinger and Wenzel Matiaske

July 2009

Federal Ministry
of Education
and Research

Working Paper Series of the Council for Social and Economic Data (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

The RatSWD Working Paper Series is edited by:

Chair of the RatSWD (2007/ 2008 Heike Solga; 2009 Gert G. Wagner)

Managing Director of the RatSWD (Denis Huschka)

Contact: Council for Social and Economic Data (RatSWD) | Mohrenstraße 58 | 10117 Berlin | office@ratswd.de

Regional Data

Gerd Groezinger and Wenzel Matiaske

Helmut-Schmidt-University / University of the Federal Armed Forces Hamburg
(matiaske[at]hsu-hh.de)

Abstract

Spatiality is an increasingly important dimension in the social sciences, as a new wave of recent publications shows. Intra-national comparisons have proved to be as fruitful as the better known inter-national analysis. Regional characteristics are found to have considerable influence on individual behaviour. This movement was fostered by methodological advances, e.g. in multi-level techniques. On the data side spatial analysis is supported by a good basic infrastructure in official and semi-official information, often provided in the way of easily usable DVDs. In addition, both scientific researchers and commercial marketing firms collect valuable information, partly on a very detailed local level of only a handful of households. However, there is ample room for improvement. Huge existing datasets (e.g. PISA E) are not open for spatial oriented scientific purposes; in many cases regional information is not sufficiently available (e.g. on criminal behaviour); systematic over-sampling in sparsely inhabited areas to allow additional regional analysis is not very common.

1. Research Questions

The topic region and social behaviour has a long tradition with different roots in social sciences. In sociology Durkheims famous book on suicide should be mentioned as one of the prior works studying the impact of regional characteristics – religiosity, urbanization and social control – on individual anomic behaviour (Durkheim 1952). The basic argument which models regional characteristics as independent variables influencing social behaviour has henceforward often been picked up. But early sociology is also known for studies, which do concentrate on the regional context embedding the social relationships within a group or community (‘Gemeinschaft’). Whytes’ well-known description of the “Street Corner Society” in Boston’s Little Italy in the tradition of the Chicago School opens up the space dimension in this understanding (Whyte 1943). In the tradition of economic theory also both argumentative traces are to be found. Marshall emphasizes the importance of regional characteristics in the shaping of industrial districts and their role as a core determinant for economic development (Marshall 1898). From the business-management point of view regional aspects are discussed as a problem of site selection. Von Thünens concentric model of land use may be read as an early version of an industrial location theory (Thünen 1826).

Both directions - regional aspects as co-founding and as a central field of social acting - still exert deep influences on the debate. The process has been precipitated by theoretical and methodological developments. On the theoretical side the discussion deals with a more clarified differentiation between macro and micro-levels of social behaviour (Alexander et al. 1987). In the words of the contemporary rational choice theory the context of action on the macro-level of social systems – regions in our context - constrain the 'logic of the situation’. Regional conditions on the macro-level influence the selection of goal-oriented actions of individual or corporative actors on the micro- (resp. meso-) level. In addition, the logic of aggregation leads back from the micro-to the macro-level of the social system and shows emergent effects there, which are not always collective goods by the ‘invisible hand’ but may also lead to situations of collective damage (Coleman 1990).

The theoretical clarifications correspond with methodological progress. Especially here to name is the development of so-called hierarchical regression models – respectively fixed and random effect models (the terminology of sociological and economic oriented methodologists differs). These regression models take the hierarchical structure of the analysis explicitly into account: behaviour or attitudes are not only explained by individual properties (micro level) but also by regional circumstances (macro level) (Snidjders and Bosker 1993). Regarding different degrees of freedom on the resp. hierarchical levels increase the reliability of the test statistic. A common property is the integration of cross level interactions. According to the needs of the analysis different estimators are available (Blien 2005). However, there is a danger of over-stretching such analysis and fall thereby victim to the 'ecological fallacy'-problem. For a proper way to model the macro-constraints of the logic of the situation, individual data and structural (regional) data must either directly mirror each other or form another link of reasoning.

Whereas this group of multi-level models primarily is predestined to analyze the macro-micro-link, there does not exist any standard model to describe the micro-macro-link. In many cases one can use a micro-economic model of market exchange to analyze the logic of aggregation, typically to study price or power effects (Braun 2008). But the assumption of more or less perfect markets does not always hold true. Therefore, a multiplicity of methods like game theory models, Markov-models and simulation studies are employed. Currently social network analysis is more and more used in the multi-level context (Wasserman/Faust 1994). Furthermore multivariate techniques developed or modified for ecological analysis, e.g. restricted or detrended correspondence analysis and other eigenvalue techniques or multi-dimsional scaling, seems to be extremely useful in the case of regional data (Leyer and Wesche 2005).

Beside the scientific dimension regional analysis was always of interest for policy makers. After WW II the compilation of German regional data has its first upswing in the late 60's and early 70's of the last century (at least in the Western part). It is connected with the new public interest in planning policies (Schäfers 1973). Scientific organisations have responded to this demand by professionalization, so that the start of many research activities dates back to this decade. E.g. the section on 'Stadt-und Regionalsoziologie' of the German Society of Sociology was officially established in 1975, the same year where another user group with regional interests, the 'Informationskreis für Raumplanung' with now over 1.500 members was founded. 1976 followed the economists' association, the 'Verein für Socialpolitik' with the establishment of a commission on 'Regionaltheorie und Regionalpolitik'.

With the deepening and enlargement of the European Union new accentuations and questions came up. Instruments like the Cohesion Fund, the Social Fund, the Regional Development Fund needed data for implantation and evaluation. Comparisons with other countries were made easier by establishing common definitions of regional units. Since 2003 a framework on the definitions of NUTS (Nomenclature des unités territoriales statistiques) was legally enacted in the EU, based on former cooperation and experiences among the national statistical offices (Brunner 2008).

Another push in interest on the regional dimension came with the German unification. Given the strong sustained differences between East and West, social sciences began to look for explanations for different development paths (e.g. Bertram et al. 2000). And the public got also more interested, which in turn did lead to many activities. A huge ‘Nationalatlas Bundesrepublik Deutschland’ project was started, where in twelve volumes/CDs a comprehensive view of the life in the German regions is offered (Leibniz-Institut für Länderkunde Div.). This done mostly on the ‘Raumordnungsregion’-level (ROR). Also on the ROR-level an online-survey was conducted which served as a basis for many comparisons in the media (Faßbender and Kluge 2006).

Of particular importance for the political arrangement is the labour market. In Germany the labour market is characterized by extensive regional disparities, especially in terms of the extent of employment and unemployment, but also in terms of the income level. The ‘Institut für Arbeitsmarktforschung’ (IAB) collects and analyses data of the labour market – employment statistics, unemployment statistics, the IAB Betriebspanel etc. – on different levels (Blien et al. 2001). IAB contains its own department of research ’Regionale Arbeitsmärkte’ and moreover carries out the ‘Regionale Forschungsnetz’ in the places of former state employment bureaus (Eckey et al. 2007).

In specialized space-and regional research, in economic research and in current business administration research - e.g. on the development of regional clusters - the region is understood as an independent object of research. In contrast in behavioural-science oriented fields of study dominates the link on macro-level – i.e. aggregate data on characterizing social environment – with behaviour, attitude and preference in micro-level (see Grözinger and Matiaske 2005; Grözinger et al. 2008 for summing up current studies). These fields of research integrate micro and macro data, which normally stem from different source of data. Below we will highlight these research facilities and will discuss characteristic aspects of space data and problems of bringing them together with individual data. The potential capacity of these datasets organized in small-scale coordinates should not be underestimated. By the possibilities of fusion of data, i.e. the link between the regionalizable datasets, which can be matched to these coordinates. Matchable datasets do not only stem from public or for scientific reasons measured dataset, but in the area of commercial research primarily also from other sources.

In contrast to that private enterprises have the main interest in regional economic respectively marketing policy. For such decisions it is not least reverted to databases, which are provided by private research facilities and business consultancies. E.g. GfK Group Nürnberg one of the biggest European providers for geo-marketing data and support analysis, planning, and evaluation of locations in Germany and abroad. Not only for practical purposes but also for research regional data based on point of sale surveys, socio-demographic and sector-specific data are of interest. Indicators for purchasing power are collected and can be analyzed at all regional levels down to individual street sections (Lochschmidt 2005). Similar data are provided by other companies. Microm e.g. calculates 'social milieus' from such data and this is also used by the SOEP-group for complement the survey information (Kueppers 2005).

2. Status Quo: Data Bases and Access

Due to the above named tradition, in the cases, where data are needed for planning purposes, a good basic infrastructure of regional information is still provided by official sources. This is partly done by the Federal Statistical Office (Destatis), often in cooperation with the Statistical Offices of the States, and a special federal research unit, the 'Bundesanstalt für Bauwesen und Raumordnung' (BBR). The BBR publishes widely used regular reports on the structure of regional differences in Germany ((Bundesamt für Bauwesen und Raumordnung 2005)) and its projections into the future (Bundesamt für Bauwesen und Raumordnung 2004).

In a hierarchical order data from Destatis and the BBR can be usually found on the following levels (numbers show the respective amount of entities):

- States (Bundesländer, in short 'BL'): 16
- Regional Planing Units (Raumordnungsregionen, in short ‚ROR'): 97
- Cities and Counties (Städte und Gemeinden, in short 'SG'): 439.

Especially three data compilations are to be mentioned. All are for scientific use easily usable since on CD/DVD, both come without user restrictions, are more or less reasonably priced (approx. 75 €) and regularly updated. In addition, there are connected webpages available where the variables are defined and maps are provided (www.raumbeobachtung.de) or even updated data are online retrievable (www.regionalstatistik.de):

- INKAR (Bundesamt für Bauwesen und Raumordnung 2007) with approx. 800 indicators
- Statistik Regional (Statistische Ämter des Bundes und der Länder 2008b) with approx.

1.100 indicators

- Statistik Lokal (Statistische Ämter des Bundes und der Länder 2008a) with more than 300 indicators.

In many cases, these datasets fulfil the interest of social researchers, looking for some regional background information. Where appropriate, differentiation along the lines of gender and migration is often included. E.g. in the case of unemployment, INKAR yields an unemployment rate of women, the absolute number, the share and the development. For foreigners the rate, share and development figures are given.

Regional information can often further be broken down to an even more detailed grid. Some of the German States are rather large in population und consist therefore of different Administrative Areas (Regierungsbezirke). Many, especially bigger cities have information broken down on Boroughs (Stadtteile / Bezirke). And on the most detailed level every municipality provides a land registry (‘Kataster’). Whereas such data must be retrieved by the regional or local administrations, detailed general information about the approx. 12.000 municipalities (Gemeinden) is easily available from a special DVD:

- Statistik lokal (Statistische Ämter des Bundes und der Länder 2007a).

However, it must be named that the respective statistical units are shaped either directly by political traditions or for planning purposes, which also rest on political boundaries. For scientific questions one has therefore to deal with huge variances in both the population and area, which can make analysis often rather difficult. E.g. the number of inhabitants, in most contexts a highly relevant information, ranges from:

- On the BL-level, the minimum is 0,7 Mill. (Bremen), the maximum 18 Mill. (Nordrhein-Westfalen)
- On the ROR-level, the minimum is less than 300.000, the maximum is Berlin with over 3 Mill.
- On the SG-level, the minimum is barely over 50.000, the maximum again Berlin with over 3 Mill.

Besides this official statistical entities, there are other classifying principles, mostly used by scientific or marketing institutions for sampling, e.g.:

- ZIP-Codes (Postleitzahlen)
- Electoral Districts (Wahlbezirke)
- Telephone Are Codes (Telefonvorwahlen)
- Labour Market Regions (Arbeitsmarktregionen)

- Licence Plates (Autokennzeichen)
- Market Cells (Haushalte).

Some of them can also be (dis)aggregated according to the needs of the user. E.g. the Zip-code has 5-digits and is hierarchical ordered. It can therefore fully be used or only by the first or first two, three or four digits.

Market Cells are the smallest unity for information sampled by marketing institutions. Although not legally determined, it is generally understanding that for privacy protection rules, in Germany every local statistical information has to be based on at least 5 Households (Mietzner 2005). But it is allowed to combine information on such clusters. On this base information won especially by consumer marketing techniques a wealth of data can be assembled to characterize a certain area due to sociological criteria.

Whereas both these above named lists rest on the principle of physical closeness, it is also possible to classify regional entities by common properties. Often used principles in the social sciences are:

- Number of inhabitants
- Income levels
- Types of urbanization.

The last one can be differentiated according to the needs and the levels of regional aggregation. The BBR offers e.g. for its data a classification of 3 general regional types of settlement, 7 types on the ROR-level and 9 on the SG-level.

A special attention deserves the SOEP ('Socio-Economic Panel'). It is by far the most-used dataset for social science questions in Germany. Registered users with appropriate data safety measures can get access to a version on the ROR-level. And directly on the premises in Berlin one can even work with a version on the BL-level (http://www.diw.de/english/soep/29012.html).

In principle, every special data set that contains information on the sampling point is a potential source for aggregation to some regional level. E.g. one may estimate the (officially not available) regional religious distribution on basis of a survey (Dülmer 2005). But the regionalized sample size must excess a critical number to provide reliable estimators (Bliese 2000).

Finally, not all data one may reasonably expect, is available on the appropriate regional level. Three examples in publicly very much debated fields: (1) the criminal statistic is not regularly and comprehensively published in such a way (Bundeskriminalamt 2008), (2) the

PISA-E-study is not made open for secondary analysis below the BL-Level, (3) the outcome of the IQ-tests of young men in connection with the military draft system is also seen as private property although it can be successfully linked to regional circumstances (Ebenrett et al. 2003).

3. Future Developments and Challenges

It is generally underestimated that the formation of regional characteristics may have a very long duration, often out spanning the usual time frame of official data. E.g. in a recent study the impact of social capital on the regional crime rate is analyzed, the authors having the opportunity to use historical data on household, population, occupation etc. as instrumental variables, starting from 1795 to 1970 (Akcomak and Weel 2008). The Netherlands Volkstellingen Archive (Dutch census) provides this data and more (http://www.volkstellingen.nl/en/). It would be an improvement if also in Germany historical regional data from different sources - church and land registers, historical reports etc. - would be properly edited and made available for quantitative analysis.

Looking over the border leads to another aspect of future research improvement. The European classification NUTS is available since several years, which makes comparative research more easy. However, this classification system is more appropriate for planning purposes than for social research. On the European level a future challenge will therefore be the development of a more detailed classification system, based on the needs of social scientists.

Generally, there is a trade off between a finely woven classification system and the data privacy. In particular, providing market cell data for geo-marketing may have negative side-effects by discriminating the inhabitants of certain areas (‘scoring’). In the long run the effect may not only lead to intra-region migration movements and a self-supporting vicious cycle of discrimination. It may also increase the public distrust of data collecting and endanger the legitimation of social science research. Furthermore, there can occur problems in the reliability of measured datasets when the data of different reference levels are brought together or methods of fusion data are applied (Zimmermann 2005).

4. Conclusions and Recommendations

The following list contains the most relevant measures to improve the infrastructure on regional data in Germany. From an organizational point of view the most relevant are:

- In addition to its publications has the BBR a huge amount of unpublished data on different regional levels on file. There should be at least a regularly updated list of such data with a proper description and a well-defined policy how to get access to the data for scientific purposes.
- The ‘Zentralarchiv’, where many of the German survey data is stored, should get extra-funds to classify all surveys according to their appropriateness for a regional analysis.
- Future surveys with the aim of being nationally representative should principally be in a way sampled that at least on the ROR-level a detailed regional analysis is also possible. Due to the different amount of inhabitants, this would need some systematic over-sampling in sparsely inhabited areas.
- The 5-houshold-entity is at the time being not formalized but could be cut by any principle. None withstanding the above named danger with illegitimate use of such information, it would be useful if the marketing firms cooperate to work out one single list of such blocks, which then can be universally used. Alternatively, the 8-household grid of the Microzensus - anyway to be renewed for the census of 2011 - could be used for such purpose.
- A concordance should then be provided, where the different levels and principles could be easily transferred upward (e.g a special ROR, ZIP-area etc. consists of certain numbers of blocks).
- Finally, the very differentiated areas of interest in regional and geographical information as well as from the scientific, administrative and commercial users and data producers leads to the recommendation of a round table where common interests could be defined. The RatSWD should initiate such a group.

References:

Akcomak, I.S. and Weel, B.T. (2008): The Impact of Social Capital on Crime: Evidence from the Netherlands, Maastricht.

Alexander, J.C./Giesen, G./Münch, R. and Smelser, N.S. (Eds.) (1987): The Micro-macro Link, Berkely.

Bertram, H./Nauck, B./Klein, T. (Eds.) (2000): Solidarität, Lebensformen und regionale Entwicklung, Opladen.

Blien, U. (2005): Die Mehrebenenanalyse regionaler Fragestellungen. In: Grözinger, G. and Matiaske, W. (Eds.): Deutschland regional. Sozialwissenschaftliche Daten im Forschungsverbund München, 133 -156.

Blien, U./Haas, A./Hirschenauer, F./Meierhofer, E./Tassinopoulos, A./Vollkommer, D. and Wolf, K. (2001): Regionale Arbeitsmarktforschung im IAB. In: Mitteilungen aus der Arbeitsmarkt- und Berufsforschung (1), 45 - 73.

Bliese, P.D. (2000): Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In: Klein, K.J. and Kozlowski, S.W. (Eds.): Multilevel Theory, Research, and Methods in Organizations San Francisco, CA, 349 - 381.

Braun, N. (2008): Sozialkapital aus Sicht der Rational Choice Soziologie. In: Matiaske, W. and Grözinger, G. (Eds.): Sozialkapital: eine (un)bequeme Kategorie (Jahrbuch Ökonomie und Gesellschaft, Band 20) Marburg, 43 - 78.

Brunner, C. (2008): European Datasets: Regional and Urban Statistics. In: Grözinger, G./Matiaske, W. and Spieß, C.K. (Eds.): Europe and its Regions. The Usage of European Regionalized Social Science Data Cambridge, 23 - 26.

Bundesamt für Bauwesen und Raumordnung (2004): Raumordnungsprognose 2020. In: Informationen zur Raumentwicklung 3/4.

Bundesamt für Bauwesen und Raumordnung (2005): Raumordnungsbericht 2005 In: Berichte 21.

Bundesamt für Bauwesen und Raumordnung (2007): INKAR - Indikatoren und Karten zur Raum- und Stadtentwicklung. Bonn.

Bundeskriminalamt (2008): Polizeiliche Kriminalstatistik 2007, Wiesbaden.

Coleman, J.S. (1990): Foundations of Social Theory.

Dülmer, H. (2005): Die Schätzung von kleinräumigen Kontextinformationen aus Umfragedaten. In: Grözinger, G. and Matiaske, W. (Eds.): Deutschland regional. Sozialwissenschaftliche Daten im Forschungsverbund München, 29 - 39.

Durkheim, E. (1952): Suicide. A Study in Sociology, London.

Ebenrett, H.J./Hansen, D. and Puzicha, K.J. (2003): Verlust von Humankapital in Regionen mit hoher Arbeitslosigkeit. In: Aus Politik und Zeitgeschichte B 6-7, 25 - 31.

Eckey, H.-F./Schwengler, B. and Türck, M. (2007): Vergleich von deutschen Arbeitsmarktregionen, Nürnberg.

Faßbender, H. and Kluge, J. (2006): Perspektive Deutschland. Was die Deutschen wirklich wollen, Berlin.

Grözinger, G. and Matiaske, W. (Eds.) (2005): Deutschland regional. Sozialwissenschaftliche Daten im Forschungsverbund, München.

Grözinger, G./Matiaske, W. and Spiess, K.C. (Eds.) (2008): Europe and its Regions. The usage of European regionalized social science data (Forthcoming), Cambridge.

Kueppers, R. (2005): MOSAIC von microm. In: Grözinger, G. and Matiaske, W. (Ed.): Deutschland regional: Sozialwissenschaftliche Daten im Forschungsverbund Müncher, Mering, 95 - 104.

Leibniz-Institut für Länderkunde (Eds.). (Div.): Der Nationalatlas Bundesrepublik Deutschland Heidelberg.

Leyer, I. and Wesche, K. (2005): Ordinationsmethoden zur Analyse ökologischer Daten. In: Grözinger, G. and Matiaske, W. (Eds.): Deutschland regional: Sozialwissenschaftliche Daten im Forschungsverbund Müncher, Mering, 157 - 170.

Lochschmidt, B. (2005): Wissen gesucht? Wissen gefunden: GfK Regionalforschung. In: Grözinger, G. and Matiaske, W. (Eds.): Deutschland regional: Sozialwissenschaftliche Daten im Forschungsverbund Müncher, Mering, 105 - 114.

Marshall, A. (1898): The Principles of Economics, London.

Mietzner, L. (2005): Anwendungsfelder für mikrogeographische Daten im Marketing In: Sokol, B. (Ed.): Living by Numbers. Leben zwischen Statistik und Wirklichkeit Düsseldorf, 38 - 52.

Schäfers, B. (Ed.) (1973): Gesellschaftliche Planung, Stuttgart.

Snidjders, T. and Bosker, R. (1993): Multilevel analysis. An introduction to basic and advanced multilevel modelling, London.

Statistische Ämter des Bundes und der Länder (2008a): Statistik lokal. Gemeindedaten für ganz Deutschland. Düsseldorf.

Statistische Ämter des Bundes und der Länder (2008b): Statistik regional. Daten für die Kreise und kreisfreien Städte Deutschlands. Düsseldorf.

Thünen, J.H. v. (1826): Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie, Hamburg.

Wasserman, S. and Faust, K. (1994): Social network analysis: Methods and Applications Cambridge.

Whyte, W.F. (1943): Street corner society: the social structure of an Italian slum, Chicago.

Zimmermann, E.J. (2005): Möglichkeiten und Grenzen der Datenfusion. In: Grözinger, G. and Matiaske, W. (Eds.): Deutschland regional. Sozialwissenschaftliche Daten im Forschungsverbund München, 171 -190.